Pitfalls of and Controversies in Cluster Randomization Trials

Allan Donner, PhD, and Neil Klar, PhD

It is now well known that standard statistical procedures become invalidated when applied to cluster randomized trials in which the unit of inference is the individual. A resulting consequence is that researchers conducting such trials are faced with a multitude of design choices, including selection of the primary unit of inference, the degree to which clusters should be matched or stratified by prognostic factors at baseline, and decisions related to cluster subsampling. Moreover, application of ethical principles developed for individually randomized trials may also require modification.

We discuss several topics related to these issues, with emphasis on the choices that must be made in the planning stages of a trial and on some potential pitfalls to be avoided. (*Am J Public Health.* 2004;94:416–422)

Cluster randomization trials, in which intact groups of individuals are randomized to receive different interventions, have been increasingly adopted by public health researchers over the past 2 decades after publication of a seminal article by Cornfield[1] and the extensive methodological developments stimulated by this article. The units of randomization for such trials are diverse, including, for example, clinics, hospitals, worksites, and entire communities. It is also well known that such trials may have substantially reduced statistical efficiency relative to trials that randomize the same number of individuals. This reduction in efficiency is a function of the variance inflation due to clustering (also known as the design effect), given by $1 + (\bar{m} - 1)\rho$, where $\bar{m}$ denotes the average cluster size and ρ is a measure of intracluster correlation, interpretable as the standard Pearson correlation between any 2 responses in the same cluster. With the additional assumption that the intracluster correlation is nonnegative, ρ may also be interpreted as the proportion of overall variation in response that can be accounted for by the between-cluster variation. The effect of variance inflation on sample size requirements is discussed in more detail later.

In practice, several attractive and methodological features of this design, including increased administrative efficiency, lessened risk of experimental contamination, and likely enhancement of subject compliance, are often perceived by public health researchers to outweigh the resulting loss in statistical precision. Moreover, in some studies, the nature of the intervention itself may dictate its application at the cluster level. This can be seen, for example, in community intervention trials designed to evaluate the effect of a health education program using the mass media.[2] Cluster randomization designs also have particular advantages when applied to vaccine field trials, since they allow both the direct and indirect effects of the intervention to be evaluated.[3] In spite of its growing popularity, however, the development of a well-accepted methodological foundation for this design has been relatively slow. Therefore, in this article, we discuss and comment on a number of selected issues relevant to the task of providing such a foundation.

We begin with a discussion of ethical considerations involving the need to obtain informed consent as it relates to the nature of the clusters randomized and to the multiple levels of consent potentially involved in this process. Development of guidelines for informed consent in cluster randomization trials has proceeded even more slowly than development of guidelines for design and analysis, which is why our discussion of this issue is fairly extensive.

The remaining sections of the article deal with issues that arise in the design stage of a trial, including the often-overlooked but fundamental decisions that must be made in choosing the primary unit of inference. Also discussed here are the advantages and disadvantages of pair matching as a design strategy, hazards related to subsampling of clusters, and instability problems related to overreliance on empirically estimated values of the parameter ρ.

ISSUES INVOLVING INFORMED CONSENT

Investigators conducting individually randomized trials of therapeutic interventions are routinely required to obtain the informed consent of study participants before their random assignment. Such a requirement not only ensures that the risks of experimentation are adequately communicated to patients but also facilitates the process of random assignment, which may at times be seen to compromise the implicit contractual relationship between patient and physician. However, the question arises as to whether such a strict analogy is required for trials in which clusters of individuals, rather than individuals themselves, are randomized to different intervention groups. This is particularly so in the case of large clusters such as entire communities, where it may be logistically difficult or even impossible to obtain informed consent in a routine manner from all targeted individuals before random assignment.

Several investigators have argued that in trials of routine health care activities these practical difficulties, combined with the relatively low risk of the intervention being assessed, may remove any need for informed consent.[4–6] However, others[7] disagree with this position and argue persuasively that attention should be given to developing special mechanisms for protecting the interests of trial participants. In this regard, Edwards et al.[8] suggested that it may be valuable from an ethical perspective to distinguish cluster randomization trials based on the level at which the intervention is offered.

For instance, if the intervention is offered at the cluster level, it is typically not possible to obtain consent before its administration (e.g., media campaigns designed to prevent drunk driving). In this case community leaders, including elected and appointed officials, could act as surrogates in providing agreement for random assignment. However, as pointed out by Strasser et al.[9] and Brody,[10] it is by no means certain when or even if

the agreement of such surrogates is sufficient, implying that other precautions should also be taken. Thus, as stated in a set of recently released guidelines for cluster randomized trials,[11] "the roles of the guardians of the patients' interests during the trial, the gatekeepers of access to patient groups, and sponsors of the research are even more important in cluster randomized trials where individuals may not have the opportunity to give informed consent to participation." Although this guideline is primarily directed toward trials of medical therapies, it would seem to apply with equal weight to prevention trials and to the evaluation of nontherapeutic interventions.

When permission from key decisionmakers associated with each cluster is needed for assigning interventions, some indication should be provided as to who these decisionmakers are and how they were identified. Some information about the consent procedure administered to individual study participants should also be provided. In particular, it would be helpful to know what opportunities, if any, existed for cluster members to avoid the inherent risks of intervention.

As a first step in developing a well-accepted set of ethical principles and norms for cluster randomization trials, editors could require all articles describing results to report having institutional review board approval and to indicate how issues of participant consent were addressed. The greater challenge is in determining what other relevant ethical features of cluster randomization trials should be reported, for example, timing of informed consent. As discussed in more detail later, it is quite common for patients to be enrolled in cluster randomization trials after random assignment, an enrollment scheme that may be seen as an example of the randomized consent design proposed by Zelen.[12] In this case, patients can only consent to provide data and may not have the ability to avoid potentially harmful effects of interventions offered at the cluster level.

These suggestions may not require the development of novel ethical criteria. Very similar suggestions were proposed in the 1991 International Guidelines for Ethical Review of Epidemiological Studies, put forward by the Council for International Organizations of Medical Sciences.[10] To promote further debate, we offer the following quotation from the Community Agreement section of these underutilized guidelines:

> When it is not possible to request informed consent from every individual to be studied, the agreement of a representative of a community or group may be sought, but the representative should be chosen according to the nature, traditions and political philosophy of the community or group. Approval given by a community representative should be consistent with general ethical principles. When investigators work with communities, they will consider communal rights and protection as they would individual rights and protection. For communities in which collective decision-making is customary, communal leaders can express the collective will. However, the refusal of individuals to participate in a study has to be respected: a leader may express agreement on behalf of a community, but an individual's refusal of personal participation is binding.[10(p225–226)]

FAILURE TO APPROPRIATELY IDENTIFY THE UNIT OF INFERENCE

The unit of inference in a cluster randomization trial could be at one of several levels, depending on how the investigators choose to frame the primary question of interest. Because the final choice in this regard will directly influence the approach taken to trial design and analysis, it should be made carefully and well in advance.

To illustrate this point, consider a randomized controlled trial evaluating the effect of safety advice provided by a general practitioner to families with young children.[13] Families assigned to the intervention group received a package consisting of standardized advice and safety leaflets, while control group families received their usual care. This intervention was ultimately shown to be effective, with families receiving the package more likely to increase their safety-related behavior (e.g., storage of medicines and cleaning materials) and to make use of designated safety equipment (e.g., stair gates). Since inferences in this trial were directed at outcomes measured at the family level only, no measurements were taken at the level of the individual. Thus, the study could be regarded, at least with respect to estimation of sample size and analysis approach, as a standard clinical trial.

Decisions about the appropriate unit of inference are more complicated when data are collected at the individual level. This complexity can be seen in a physician-randomized trial in which the goal was to reduce the total number of tests ordered per clinical problem.[14] From this perspective, it followed that any variables measured at the patient level were not of particular interest; what was most important to the investigator was the effect of the intervention on outcomes that were aggregated at the physician level.

However, the ultimate goal of most health research studies is to reduce morbidity and mortality, with inferences directed at the level of the individual.[15] This was almost certainly the case, for example, in the community randomized trial reported by West et al.[16] that examined the efficacy of vitamin A in reducing mortality among preschool children in Nepal.

Additional insight into this issue can be gained by considering the RE-AIM framework developed by Glasgow et al.[17] to evaluate the impact of health promotion interventions. The 5 dimensions of RE-AIM (reach, efficacy or effectiveness, adoption, implementation, and maintenance) were selected to reflect the different levels at which interventions influence public health while balancing concerns for internal and external validity. For example, the efficacy of an intervention is typically directed at the level of the individual, while improvements in the implementation of an intervention already shown to be efficacious are usually directed at the cluster level. It is also interesting to note the parallel (and seemingly independent) focus on implementation research recently seen among both primary care[18] and public health[17] researchers.

The examples just discussed may also be used to show how the selected unit of inference affects the choice of randomization unit. Thus, if the unit of inference is at the level of the individual, the investigators may have considerable flexibility in selecting the unit of randomization. This is reflected, for example, in a meta-analysis synthesizing the results from 12 trials investigating the effect of vitamin A supplementation on child mortality.[19] In 4 hospital-based trials individual children served as the unit of allocation, while in the 8

community-based trials the allocation units included households, neighborhoods, and entire communities.[16] Considerably less flexibility in this regard will exist if the unit of inference is explicitly intended to be at a higher level, as in the study just referred to that aimed to reduce the number of tests ordered by a physician for a given clinical problem.[14]

The choice of unit of inference will also frame the approach taken to the statistical analysis. For example, in the vitamin A trial reported by West et al.,[16] the investigators displayed a between-group comparison with respect to selected individual-level baseline characteristics, each of which was a candidate for subsequent statistical adjustment. On the other hand, Verstappen et al.[14] omitted any consideration of patient-level characteristics in their analysis, since inferences in their trial were exclusively directed at the level of the physician.

Decisions concerning the unit of inference are also affected by secondary analyses that might be planned. The reason is that certain predictor variables may be conceived to exist at either the individual or the cluster level, raising interpretational issues related to the well-known ecological fallacy. For example, the proportion of patients served by a practice who were at least 65 years of age was a predictor variable considered by Verstappen et al.[14] This predictor variable may have a different association with a physician's performance in ordering tests than the age of an individual patient. Note that the ecological fallacy cannot arise when inferences are constructed about the effect of intervention, since the assigned intervention is shared by all cluster members.

Finally, selection of the unit of inference may also directly affect interpretation of the study results. Thus, a small effect size at the community level may have much more practical significance than an effect of comparable size at the individual level. That is, the threshold for determining the magnitude of an effect necessary to have public health significance may be quite different than the threshold for "clinical significance,"[20] a consequence of the prevention paradox,[21] according to which (in the context of the RE-AIM framework) "[l]ow-intensity interventions that are less efficacious but can be delivered to large numbers of people may have a more pervasive impact."[17(p1322)]

The issue of choosing the unit of inference is sometimes referred to as the "unit of analysis problem." We believe that this phrase can be misleading, since it confuses the choice of analytic unit with the need to account for clustering. Similarly, statements sometimes seen in the literature to the effect that "analysis by individual" is incorrect for cluster randomization trials or that the "allocation unit should be the unit of analysis" are also misleading. In general, an analysis at the individual level that properly accounts for the effect of clustering is equivalent to an appropriately weighted cluster-level analysis. Thus, the issue of fundamental importance in this context is best referred to as the unit of inference, rather than the unit of analysis.

ISSUES INVOLVING MATCHING AND STRATIFICATION

Overmatching

Pair matching is a long-standing and popular strategy in epidemiological research. This is partly because matching provides a kind of "face validity" to the design that makes it clear that potentially important confounders are taken into account. Pair matching may also lead to a gain in power relative to an unmatched design, provided that an effective set of matching factors can be identified.

Matching by baseline factors such as cluster size and geographic area is a particularly common strategy in the design of trials randomizing a relatively small number of clusters, where it may be feared that failure to match could lead to poor randomization. In spite of these advantages, however, the actual benefits of matching in practice will not be realized unless several conditions are satisfied, conditions that may be difficult to achieve in practice.

Many cluster randomization trials reported in the literature have recruited and enrolled a fairly large number of matched pairs. For example, Ray et al.[22] reported the results of a physician-based pair-matched trial investigating an educational program designed to reduce use of nonsteroidal anti-inflammatory drugs (NSAIDs) among community-dwelling elderly people. The 220 eligible physicians were stratified, according to number of elderly NSAID users, into 110 strata and randomly assigned either to an educational program or to a control group. However, recruiting such a large number of pairs is worthwhile only to the extent that these pairs represent different levels of baseline risk; otherwise, there is no statistical advantage in terms of increased power. It would seem more sensible to pool those clusters having nondistinguishable baseline risks into a single stratum, creating, for example, 5 strata of size 44 instead of 110 pairs (strata) of size 2. This not only would increase the number of degrees of freedom available for estimation of error variation but would also allow much more flexibility in the data analysis.

The latter advantage arises because the intracluster correlation coefficient (ICC) cannot be directly estimated from a pair-matched design owing to the confounding between the effect of the intervention and the natural variation that exists between 2 clusters in a matched pair, even in the absence of intervention. This inevitably hampers the planning of future studies involving the same outcome variables and unit of randomization, because the size of the required sample is very sensitive to the magnitude of ρ. Furthermore, since the test of intervention effect for a matched-pair design must be based on error variation computed among rather than within pairs, application of regression modeling procedures such as generalized estimating equations and mixed-effect regression analysis is no longer routine.[23]

Feng et al.[24] raised a potential pitfall that may occur in the analysis of data arising from a stratified cluster randomization trial. In particular, they stated that the validity of permutation tests may be questionable if there is imbalance in cluster size or in the number of clusters per intervention group within strata, leading to an imbalance in overall group sizes. However, since one typically stratifies by cluster size in such designs, the theoretical concerns raised by Feng et al.[24] (as identified earlier by Gail et al.[25]) should have minimal impact in practice. Moreover, it should be noted that Gail et al.[25] limited attention to completely randomized and pair-matched designs. It is therefore unclear as to the actual degree of imbalance needed to seriously dis-

tort the properties of a stratified permutation test. It is also apparent that these difficulties can be avoided by constraining randomization within a stratum to ensure that balance is achieved between the number of clusters assigned to each intervention.[26]

In the case of trials involving a small number of clusters, say 10 pairs or less, the loss of degrees of freedom resulting from pair matching becomes a particularly critical factor. Detailed investigation of this problem led Martin et al. to conclude that "for small studies, it is unlikely that effective matching would be possible" and that "matching may be overused as a design tool."[27(p336–337)] An additional disadvantage in small trials is that loss to follow-up of a single member of a pair may lead to a serious loss of power if the matching is preserved, since then essentially both members of the pair are lost from the analysis.

It has been suggested by Diehr et al.[28] that breaking pair matches and treating the design as completely randomized could be a useful strategy for dealing with ineffective matching. For example, the matching might be broken if the observed matching correlation falls below a certain threshold. However, given the data-driven nature of this strategy, more research is needed on its statistical properties, particularly its effects on overall type I error, before it can be recommended for routine application.

Are Strata Fixed or Random?

An important aspect of fitting statistical models is the characterization of predictor variables as either fixed or random. Following Kleinbaum et al.,[29] a factor should be modeled as fixed if the selected levels are either the only possible ones or the only ones of scientific interest. Conversely, a variable should be modeled as random if the levels constitute a random sample, or at least a representative sample, from some larger population.

It is well known that the decision to model a factor as random tends to increase variability, leading to a loss in statistical power. For example, statistical inferences in a cluster randomization trial are generally not restricted to the selected clusters; rather, they more accurately reflect the variability of the intervention effect under the assumption that it will be applied to a new sample of clusters. Therefore, the cluster effects in such studies are inevitably regarded as random, with the subsequent loss of efficiency associated with this design. On the other hand, the effect of intervention is invariably modeled as fixed, since the study is designed to compare only the selected experimental and control conditions.

It may not always be so obvious as to whether a factor should be modeled as fixed or random. Thus, a challenge for randomized trials in general[30] and for cluster randomization trials in particular[23,24] is to make this choice in the context of the principal trial objectives and the nature of the sampling scheme. Further insight into this issue may be obtained by applying the definition given by Kleinbaum et al.[29] to several cluster randomization trials reported in the literature.

We first consider a family randomized trial reported by Farr et al.[31] This trial examined the effect of virucidal nasal tissues, as compared with placebo tissues, on the prevention of upper respiratory infections. The first stratum consisted of families having only 1 or 2 children, while the second stratum consisted of families with 3 or more children. Since the 2 levels include all possible families having at least 1 child, these strata are naturally regarded as fixed. Moreover, the study power would have been unnecessarily compromised had family size erroneously been modeled as random.

The World Health Organization's antenatal care randomized trial for evaluation of a new model of routine antenatal care provides a second example.[32] In this trial, clinics in Argentina, Cuba, Saudi Arabia, and Thailand were randomly assigned to provide either an experimental or a standard antenatal care program to expectant mothers. The 53 participating clinics were stratified both by country and by a measure of clinic size (small, medium, large) before random assignment.

In this trial, clinic size is most properly thought of as fixed, since it encompassed all possible values of this variable. However, the decision as to whether clinic location should be modeled as fixed or random is more difficult. As noted by Kleinbaum et al., location may be modeled as either fixed or random "depending on whether a set of specific sites or a larger geographic universe is to be considered."[29(p425)] While investigators in the antenatal trial were interested in extrapolating results beyond the 4 sites, these countries were a highly selected group that had the infrastructure necessary to run the trial and were willing to participate. Consequently, strata were modeled as fixed factors, implying that extrapolation to other settings must be based on clinical experience and judgment.

As a final example, consider the British Family Heart Study,[33] a pair-matched cluster randomization trial designed to evaluate the effect of a nurse-led cardiovascular screening and intervention program on reducing selected risk factors such as cholesterol level. Thirteen towns were selected by the investigators "to give a geographic spread across Britain."[33(p2064)] Within each town, pairs of medical practices were then chosen on the basis of their similar sociodemographic characteristics. In this trial, as in many others, it seemed a reasonable decision to model the pairs as random, since this choice reflected the investigators' goal of obtaining a representative sample of communities.

Interestingly, as noted by Feng et al.,[24] the analyses appropriate to pair-matched designs are identical regardless of whether strata are modeled as fixed or random. This equivalence results because the variance of the estimated intervention effect must be obtained through the use of between-stratum information (as described in the previous section). However, the situation is quite different in stratified designs involving several clusters per stratum. Thus, suppose the British Family Heart Study had assigned each intervention to at least 2 practices within each of the participating towns. It would then be possible for this trial to directly measure the degree of between-practice variation, thus separating such variation from the variation between towns. Note, however, that the results of such analyses would then differ according to whether towns were regarded as fixed or random.

ISSUES INVOLVING SUBSAMPLING

Misconceptions Concerning the Influence of Subsample Size on Study Power

As a technique for reducing the cost of data collection, some cluster randomization

trials are designed to enroll only a sample of eligible cluster members, that is, to adopt a subsampling strategy. This is often quite reasonable, since it is known that increasing the number of clusters enrolled in a trial has a greater impact on statistical power than increasing the number of participants sampled per cluster. However, the actual impact of this general result on trial design has occasionally been overstated in the literature.

For example, consider a cluster randomized trial with 2 intervention groups in which the study outcome is assumed to be normally distributed with a common variance. Suppose also that clusters are randomly assigned to intervention groups in the absence of any stratification or matching. Then the number of clusters required per group can be obtained from standard sample size formulas (e.g., see chapter 5 of Donner and Klar[34]).

Now suppose that results from earlier studies indicate that the value of ρ is approximately 0.001, typical of that seen in community intervention trials. Then increasing the subsample size from 100 to 1000 may result in more than a 5-fold reduction in the number of clusters required per group. To illustrate this, suppose 10 clusters are required per group when a subsample of 1000 participants per cluster is enrolled. This requirement is easily shown to escalate to 55 clusters per group if only 100 participants are subsampled in each cluster. In this case, an increase in subsample size is seen to have dramatic effects on power.

In general, it can be shown that the effect of increasing subsample size on study power is very sensitive to the underlying value of ρ. For a fixed number of clusters per group, the largest gains in power will be obtained when the subsample size increases from 1 to about $m=1/\rho$. Further increases will tend to have only a modest effect on power. For example, if $\rho=0.001$ in the preceding example, an increase in subsample size from 1000 to 10 000 would reduce the required number of clusters per group from 10 to 6, an effect that is relatively modest. We conclude that sensitivity analyses examining the effect of varying the number of participants sampled per cluster along with the anticipated degree of intracluster correlation remain the most reliable approach for estimating overall trial size.

Subsampling Bias

Now assume that the decision has been made concerning the size of the subsamples to be selected from each cluster and consider the mechanism by which the subsampling is to be conducted. Unbiased estimates of the effect of intervention can be assured only if analyses are based on data from all eligible cluster members or, alternatively, from a random sample of these study participants.[35] Moreover, random subsampling is easiest to implement when the respondents are selected before randomization of their cluster takes place and adequate precautions are taken, where necessary, to conceal the intervention assignment. Lack of concealment may be a source of selection bias, for example, when participants' allocation can be predicted from their home address.[36] In this case, the problem is best addressed by developing objective measures of eligibility and ensuring that the allocation is carried out by someone who is independent of the trial.[11]

When subsampling is done after randomization, selection bias may result from one of several sources. For example, consider a community intervention trial in which the statistical analysis is confined to those participants within the community who visit a clinic or health facility during the trial follow-up period. In the absence of both investigator and participant blindness, a serious risk of "participation bias" may arise, depending on the nature of the intervention. Evidence of such bias may be revealed by a substantially larger participation rate among experimental group than control group participants. Similar selection bias problems may arise in trials randomizing general practices in which physicians are asked to identify as well as to treat selected patients in a "case-finding" study. If physicians in the experimental arm are more diligent in seeking out patients than physicians in the control arm or, because of a greater level of enthusiasm, tend to identify patients for treatment who are less ill, then the measured effect of intervention could be subject to bias.

Participant blindness to intervention assignment, which is generally considered one of the pillars of validity for comparative studies, is often difficult to arrange in cluster randomization trials. As a result, refusal rates might well vary by intervention group when, for example, control participants feel psychologically disadvantaged by not being offered a new intervention. This is yet another potential source of bias, one that can be addressed at least partially by collecting and comparing relevant baseline information on the nonparticipating individuals in each group. At a minimum, investigators are urged to comply with the CONSORT (Consolidated Standards of Reporting Trials) statement,[37] which requires that the number of individuals refusing to participate be reported for each intervention group.

In some trials, the study outcome is based on routinely collected statistics (e.g., mortality data) and so will be measurable for all eligible participants. The potential for bias in estimating the effect of intervention due to differential participation across intervention groups is then a problem of compliance. Methods developed to estimate the efficacy of intervention for such trials[38,39] should, if adopted, be seen only as supplementing rather than replacing an analysis by intention to treat.

ASSESSING THE VALUE OF THE ICC FROM SMALL STUDIES

A frequently occurring difficulty in the planning of trial size is that the value of the intracluster correlation ρ may well have been estimated from a previous study enrolling a relatively small number of clusters. Given the large standard error associated with sample estimates of ρ, it would be dangerous to overestimate the stability of a sample estimate obtained from a trial involving less than about 40 clusters.

To illustrate this point, consider trials in primary care, where values of ρ range from less than 0.01 to about 0.05.[40] Assuming that the outcome variable of interest is quantitative, one can calculate the expected upper limit of a 2-sided 95% confidence interval for ρ using an approach described by Donner.[41] Consider, for example, a trial randomizing a total of 10 general practices, each enrolling 1000 patients. If $\rho=0.05$, then the expected upper 95% confidence interval for this parameter is given by 0.15. Increasing the total number of practices to 20 yields much less uncertainty,

with the expected value of the upper confidence interval now given by 0.10.

Note also that if the outcome variable of interest is dichotomous rather than quantitative, yet more uncertainty is introduced; in this case, the expected value of the upper confidence interval for ρ may be much higher. [42] Further evidence that considerable sample sizes are required to estimate the ICC with reasonable precision has been provided by Ukoumunne.[43]

These results demonstrate the importance of performing a sensitivity analysis that explores the impact of various values of ρ on the final sample size calculations and then being as conservative as circumstances permit. Complicating the matter further is that the value of ρ obtained from previous studies will depend on the specific covariates considered in the analysis and the level of stratification used in the design.

The inherent instability associated with sample estimates of ρ, coupled with the proximity of these estimates to zero, has occasionally tempted investigators to test the null hypothesis H_0: ρ=0, followed by interpretation of a nonsignificant result as evidence that clustering effects may be ignored. However, this practice should be discouraged on the grounds that the power of such tests to detect small but substantively important values of ρ will usually be abysmal.[44]

CONCLUSIONS

The past 5 years have seen a proliferation of literature dealing with methodological challenges associated with the design and analysis of cluster randomized trials. This literature includes 2 books,[34,45] special issues of leading journals dedicated to this topic,[46,47] and articles dealing with specific application areas, such as community intervention trials,[48] interventions against infectious diseases,[49] and family practice research.[50] The CONSORT statement[37] has also been extended to encompass cluster randomized trials,[51] and methodological guidelines have begun to be published and disseminated by national granting agencies.[11]

In spite of these rapid developments, there is still a considerable need for expository papers that bring these results to the attention of public health researchers. The main purpose of this article has been to highlight certain points that we have found are still not well appreciated by trial investigators, such as the choice of unit of inference, or have yet to receive sufficient attention, such as the role and implementation of informed consent guidelines. For reasons of space, we have also focused here on issues arising in the design of cluster randomized trials. Issues arising at the analysis stage of a trial, such as the choice between population-averaged and cluster-specific approaches[52] and methods to be used in multilevel analyses,[53] deserve further attention as well. ■

About the Authors

Allan Donner is with the Department of Epidemiology and Biostatistics, University of Western Ontario, London. Neil Klar is with the Division of Preventive Oncology, Cancer Care Ontario, Toronto.

Requests for reprints should be sent to Allan Donner, PhD, Department of Epidemiology and Biostatistics, University of Western Ontario, Room K201, Kresge Building, London, Ontario, Canada N6A 5C1 (e-mail: donner@biostats.uwo.ca).

This article was accepted August 27, 2003.

Contributors

Both authors helped to conceptualize ideas, interpret findings, and review drafts of the article.

Acknowledgment

Our work was partially supported by grants from the Natural Sciences and Engineering Council of Canada.

References

1. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol.* 1978;108:100–102.

2. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): 1. Cohort results from a four-year community intervention. *Am J Public Health.* 1995;85:183–192.

3. Halloran ME, Longini IM Jr, Struchiner CJ. Design and interpretation of vaccine field studies. *Epidemiol Rev.* 1999;21:73–88.

4. Goldberg HI, McGough H. The ethics of ongoing randomization trials: investigation among intimates. *Med Care.* 1991;29(suppl 7):JS41–JS48.

5. Henderson WG, Demakis J, Fihn SD, et al. Cooperative studies in health services research in the Department of Veterans Affairs. *Control Clin Trials.* 1998; 19:134–148.

6. Winkens RA, Knottnerus JA, Kester AD, et al. Fitting a routine health-care activity into a randomized trial: an experiment possible without informed consent? *J Clin Epidemiol.* 1997;50:435–439.

7. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? *Stat Med.* 2001;20:473–488.

8. Edwards SJL, Braunholtz DA, Lilford RJ, et al. Ethical issues in the design and conduct of cluster randomised controlled trials. *BMJ.* 1999;318:1407–1409.

9. Strasser T, Jeanneret O, Raymond L. Ethical aspects of prevention trials. In: Doxiadis S, ed. *Ethical Dilemmas in Health Promotion.* New York, NY: John Wiley & Sons Inc; 1987:183–193.

10. Brody BA. *The Ethics of Biomedical Research: An International Perspective.* New York, NY: Oxford University Press Inc; 1998.

11. *Cluster Randomised Trials: Methodological and Ethical Considerations.* London, England: Medical Research Council; 2002.

12. Zelen M. Randomized consent designs for clinical trials: an update. *Stat Med.* 1990;9:645–656.

13. Clamp M, Kendrick D. A randomised controlled trial of general practitioner safety advice for families with children under 5 years. *BMJ.* 1998;316:1576–1579.

14. Verstappen WHJM, van der Weijden T, Sijbrandij J, et al. Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial. *JAMA.* 2003;289:2407–2412.

15. Feldman HA. Selecting endpoint variables for a community intervention trial. *Ann Epidemiol.* 1997; 7(suppl):S78–S88.

16. West KP, Pokhrel RP, Katz J, et al. Efficacy of vitamin A in reducing preschool child mortality in Nepal. *Lancet.* 1991;338:67–71.

17. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health.* 1999;89: 1322–1327.

18. Foy R, Eccles M, Grimshaw J. Why does primary care need more implementation research? *Fam Pract.* 2001;18:353–355.

19. Fawzi WW, Chalmers TC, Herrera MG, et al. Vitamin A supplementation and child mortality. *JAMA.* 1993;269:898–903.

20. Sorensen G, Emmons K, Hunt MK, et al. Implications of the results of community intervention trials. *Annu Rev Public Health.* 1998;19:379–416.

21. Rose G. *The Strategy of Preventive Medicine.* Oxford, England: Oxford University Press Inc; 1992.

22. Ray WA, Stein CM, Byrd V, et al. Educational program for physicians to reduce use of non-steroidal anti-inflammatory drugs among community-dwelling elderly persons. *Med Care.* 2001;39:425–435.

23. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Stat Med.* 1997;16:1753–1764.

24. Feng Z, Diehr P, Peterson A, et al. Selected statistical issues in group randomised trials. *Annu Rev Public Health.* 2001;22:167–187.

25. Gail MH, Mark SD, Carroll RJ, et al. On design considerations and randomization-based inference for community intervention trials. *Stat Med.* 1996;15: 1069–1092.

26. Piantadosi S. *Clinical Trials: A Methodologic Perspective.* New York, NY: John Wiley & Sons Inc; 1997.

27. Martin DC, Diehr P, Perrin EB, et al. The effect of matching on the power of randomized community intervention studies. *Stat Med.* 1993;12:329–338.

28. Diehr P, Martin DC, Koepsell T, et al. Breaking the matches in a paired *t*-test for community interventions when the number of pairs is small. *Stat Med.* 1995;14:1491–1504.

29. Kleinbaum DG, Kupper LL, Muller KE, et al. *Applied Regression Analysis and Other Multivariate Methods.* New York, NY: Duxbury Press; 1998.

30. Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials.* 1986;7:267–275.

31. Farr BM, Hendley JO, Kaiser DL, et al. Two randomized controlled trials of virucidal nasal tissues in the prevention of natural upper respiratory infections. *Am J Epidemiol.* 1988;128:1162–1172.

32. Villar J, Ba'aqeel H, Piaggio G, et al. WHO antenatal care randomised trial for the evaluation of a new model of routine antenatal care. *Lancet.* 2001;357: 1551–1564.

33. Thompson SG, Pyke SD, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Stat Med.* 1997; 16:2063–2079.

34. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research.* London, England: Arnold; 2000.

35. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ.* 2001;322:355–357.

36. Jordhoy MS, Fayers PM, Ahlner-Elmqvist M, et al. Lack of concealment may lead to selection bias in cluster randomised and trials of palliative care. *Palliat Med.* 2002;16:43–49.

37. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet.* 2001;357: 1191–1194.

38. Albert JM. Estimating efficacy in clinical trials with clustered binary responses. *Stat Med.* 2002;21: 649–661.

39. Frangakis CE, Rubin DB, Zhou XH. Clustered encouragement designs with individual non-compliance: Bayesian inference with randomization, and application to advance forms. *Biostatistics.* 2002;3:147–164.

40. Smeeth L, Sui-Woon NE. Intraclass correlation coefficients for cluster randomised trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Control Clin Trials.* 2002;23:409–421.

41. Donner A. Sample size requirements for interval estimation of the intraclass kappa statistic. *Commun Stat Simulation Computation.* 1999;28:415–428.

42. Donner A, Eliasziw M. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. *Biometrics.* 1994;50:550–555.

43. Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomised trials. *Stat Med.* 2002;21: 3757–3774.

44. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol.* 1996;49:435–439.

45. Murray DM. *Design and Analysis of Group-Randomised Trials.* New York, NY: Oxford University Press Inc; 1998.

46. Campbell MJ, Donner A, Elbourne DR, eds. Design and analysis of cluster randomised trials. *Stat Med.* 2001;20(theme issue):329–496.

47. Donner A, Klar N, eds. Cluster randomization trials. *Stat Methods Med Res.* 2000;9(theme issue):79–179.

48. Koepsell TD, Wagner EH, Cheadle AC, et al. Selected methodological issues in evaluating community-based health promotion programs. *Annu Rev Public Health.* 1992;13:13–57.

49. Hayes RJ, Alexander ND, Bennett S, et al. Design and analysis issues in cluster-randomised trials of interventions against infectious diseases. *Stat Methods Med Res.* 2000;9:95–116.

50. Campbell MJ. Cluster randomised trials in general (family) practice research. *Stat Methods Med Res.* 2000; 9:81–94.

51. Elbourne DR, Campbell MK. Extending the CONSORT statement to cluster randomization trials: for discussion. *Stat Med.* 2001;20:489–496.

52. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res.* 1992;1:249–273.

53. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analyses. *Am J Public Health.* 1998;88:216–222.